

Statistical methodology for ANOVA-Like Differential Expression (ALDEx) analysis

The number of reads obtained for given gene in an RNA-seq dataset convey information about the relative contribution of that gene to the total number of reads, but conveys no information about the absolute expression level of the gene. Furthermore, RNA-seq datasets contain many genes with zero read counts (e.g. [1-4]) due to sparse sampling. The method described below explicitly accounts for the proportional nature of the data and the probability that genes with 0 read counts actually represent non-expressed genes as opposed to those sampled with insufficient sequencing depth. The overall approach was to estimate the technical variance for each gene in each sample, and to identify as differential those genes where the separation between conditions was substantially larger than could be explained by technical variance.

We assumed that the read count, n , for each gene was sampled from a multinomial Poisson process of rate λ , i.e., $\{[n_1, n_2, \dots] | n\} \sim \text{Multinomial}(p_1, p_2, \dots | n)$ where the proportion for each gene can be represented by $p_i = \lambda_i / \sum_k \lambda_k$. The set of proportions p_i was estimated directly from the set of counts n_i using standard Bayesian techniques [5]: the posterior distribution of $[p_1, p_2, \dots]$ was inferred as the product of the multinomial likelihood with a Dirichlet($\frac{1}{2}, \frac{1}{2}, \dots$) prior. We generated 2048 instances from this distribution to estimate the technical variance inherent in tag-sequencing data [6]. Dirichlet-distributed random variates have the property that none of the inferred proportions are exactly 0, and are constrained to enforce conservation of proportion such that $\sum_k p_k = 1$. The distribution also encapsulates the precision of estimating the proportion as a function of the total number of

read counts. All resulting operations were performed on the set of 2048 instances drawn from the Dirichlet samples rather than summary statistics because of the extreme non-normality that can arise when the read counts are small or 0.

Values from proportional distributions cannot be independent because these distributions are constrained to have a constant sum. Therefore, an increase in one or more proportions induces a concomitant decrease in one or more other proportions and *vice versa* [7]. Fortunately, the isometric log transformation [8] has been developed to transform the proportional values into linearly independent components. The procedure for a set of m proportions is simply to take component-wise logarithms and then to subtract the mean value of the resulting components. This results in the transformation of $p \mapsto q$, i.e.,

$$q_i = \log(p_i) - \frac{1}{m} \sum_{k=1}^m \log(p_k).$$

When base 2 logarithms are used, the adjusted log-proportion

values q_i correspond exactly to “fold-based” abundances. This transformation has two major effects [8]. First, it removes the possibly-large multivariate statistical bias introduced by the log-transformation. Second, the covariance induced in the adjusted log-proportions q_i can be shown to be inversely proportional to the number of genes considered. For high-throughput data where thousands of genes are simultaneously considered, this induced covariance approaches zero. Although still not independent since $\sum_k q_k = 0$, carefully-constructed analyses can exploit this near-zero covariance in order to simplify numerical computations.

In the description of the method that follows, the individual genes using i , the different conditions using j and the different replicates of a condition using k . We use medians rather than means throughout to mitigate the effects of heavy tails and skewness that are prevalent in the distributions obtained from genes with low or 0 read-counts. Recall that we are dealing with Dirichlet-distributed random variates rather than point estimates for each gene in each sample in each condition. Similarly, the set of q_{ijk} are random variables as they are simple functions of the those Dirichlet-distributed variables. We now calculate the following random variable mixtures.

- The within-condition mixture distribution

$$W(i,j) = \sum_{k=1}^{K_j} q_{ijk}$$

- The absolute difference for gene i between-conditions j and j'

$$\Delta_A(i,j,j') = W(i,j) - W(i,j')$$

- The between sample, within-condition difference

$$\Delta_W(i,j) = \max_{k \neq k'} |q_{ijk} - q_{ijk'}|$$

- The relative effect-size between conditions

$$\Delta_R(i,j,j') = \Delta_A(i,j,j') / \max \{\Delta_W(i,j), \Delta_W(i,j')\}$$

The distribution of $W(i,j)$ is termed the “within-condition distribution” of gene i within condition j . Note that the “max” operator in the definition of Δ_W makes it a conservative representative of the pooled within-condition variance across conditions. The distribution of Δ_A is termed the “between-condition” distribution. Since the distributions of Δ_A , Δ_W , and Δ_R are estimated from multiple independent Monte Carlo realizations of their underlying Dirichlet-distributed proportions for all genes i simultaneously, they intrinsically obey the requisite $\sum_i q_{ijk}=0$ constraint. These distributions are thus unaffected by differing total-read counts.

The distributions of these random variables can be fairly broad or possess heavy tails especially when the corresponding read counts are low or 0. Thus, we summarize them by their quantiles with the notation: Δ_A^{50} , and Δ_A^{1-99} denoting the median, and 1–99% quantiles for Δ_A , and so on for the others. We further identify the quantile of zero in the distribution of Δ_A as $\Delta_A^{Q_0}$. We use a symmetric variant of $\Delta_A^{Q_0}$ such that $\Delta_A^{Q_0} \leq 0.5$. note that throughout we used the full distribution of Dirichlet-distributed values. A graphical depiction of our approximate-ANOVA is depicted in Figure S9.

Differential expression was determined using the thresholds for $\Delta_R \geq 2$ and $\Delta_A^{Q_0} \leq 0.01$. A value of $\Delta_R \geq 2.0$ means that the difference between conditions is at least twice widest within-condition variance, while $\Delta_A^{Q_0} \leq 0.01$ implies fewer than 1% of the values in the $W(i,j)$ and $W(i,j')$ distributions overlap. Changing threshold values corresponds to adjusting the

stringency of a hypothesis test, with concomitant trade-off in resultant Type-I and Type-II errors.

ALDEx is available from: <http://code.google.com/p/aldex>

References

1. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Research* 2008, **18**(9):1509-1517.
2. Polymenakou PN, Lampadariou N, Mandalakis M, Tselepidis A: **Phylogenetic diversity of sediment bacteria from the southern Cretan margin, Eastern Mediterranean Sea.** *Syst Appl Microbiol* 2009, **32**(1):17-26.
3. Crawford JE, Guelbeogo WM, Sanou A, Traore A, Vernick KD, Sagnon N, Lazzaro BP: **De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-seq technology.** *PLoS One* 2010, **5**(12):e14202.
4. Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR: **RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture.** *ISME J* 2011, **5**(7):1133-1142.
5. Jaynes ET, Bretthorst GL: *Probability Theory: the Logic of Science:* Cambridge, UK: Cambridge University Press; 2003.

6. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, Sodergren E, Weinstock G, Shannon WD: **Hypothesis testing and power calculations for taxonomic-based human microbiome data.** *PLoS One* 2012, **7**(12):e52078.
7. Pearson K: **Mathematical Contributions to the Theory of Evolution.--On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs.** *Proceedings of the Royal Society of London* 1896, **60**:489-498.
8. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C: **Isometric Logratio Transformations for Compositional Data Analysis.** *Math Geol* 2003, **35**(3):279-300.

Supporting Figure Legends and Table Header Descriptions

Figure S1

Agilent 2100 Bioanalyzer electropherograms showing bacterial RNA samples before and after rRNA depletion by MICROBExpress. The large 16S and 23S peaks that make up much of the total RNA sample have been reduced by >99% after a single round of MICROBExpress. The bioanalyzer output for sample N30 after treatment was not available, but rRNA depletion was validated before sequencing. Note the difference in the relative fluorescence scale (y-axis) between plots.

Figure S2

Heatmap representation of the mean \log_2 expression of reference sequences (A) and the mean \log_2 expression of the reference sequences grouped by SEED subsystem 4 (B) for each sample. The mean \log_2 expression values were computed into a distance matrix using the 'manhattan' method in R. The values of the distance matrix are represented in the heatmaps, and pairs of samples closer to 0 are more similar in \log_2 expression. Grouping reference sequences by function (B) results in a tighter clustering of samples within a condition (shorter branch lengths and lower distance values).

Figure S3

Select SEED subsystem strip plots showing enriched functions for *L. iners* differential expression analysis. Each point is a refseq belonging to *L. iners* and points are colored red (BV) or blue (N) if significantly differential in their respective conditions. Points are plotted

by the median \log_2 difference in expression between conditions: the value is positive if the refseq is more highly expressed in BV, and negative if it is more highly expressed in N.

Figure S4

COG (Clusters of Orthologous Groups of proteins) function distribution for *L. iners* refseqs per sample. The first three plots show the fraction of reads belonging to each COG category for the *L. iners* pan-transcriptome (“all mapped reads”), the top 10% expressed refseqs, and the total differential reads. The final plot shows the “binned” number of differential genes in each COG category. The color legend below also applies to Figure 2 in the main manuscript

Figure S5

Strip plot summary of absolute expression differences for subsys4 functions binned by SEED functional categories. Differential expression analysis was performed using ALDEX for refseqs grouped by SEED subsystem 4. Each point represents one subsys4 functional group, and points are plotted by SEED subsystem level 1 (A), subsystem level 2 (B), or subsystem level 3 (C) categories. Points are colored blue (non-BV) or red (BV) if significantly differentially expressed between conditions.

Figure S6

Heatmap summary of the transcriptional effort (expression) by the bacterial community per SEED functional category. The expression of each subsys4 function as calculated by ALDEX (relative to the average of all subsys4 functions) was summed and divided by the

number of subsys4 per SEED subsystem level 1 (A), subsystem level 2 (B), or subsystem level 3 (C) categories.

Figures S7 and S8

Function/fraction plots showing the taxonomic assignment of the reads belonging to significantly differential functions. The paired barplots represent the two samples for one function. Shown are significantly differential functions of subsys level 3 for N condition (Figure S7A) and BV (Figure S7B). Significantly differential functions for KEGG analysis at pathway level (hier3) are shown for N condition (Figure S8A) and BV (Figure S8B). The colored legend for taxon assignments is presented in Figure 1 of the main manuscript text. A table lookup for each set of plots lists the functions plotted for each category (Tables S10 to S13).

Figure S9

Outline of the statistical analysis of meta-transcriptome data (AF et al, PLoS). Step 1, reads are binned according to SEED, KEGG or COG function. In the example, there are two samples from two conditions B (red) and N (blue). Step 2, read values are converted to proportions by drawing from the Dirichlet distribution, 2048 instances are collected to estimate technical variance due to sampling and sequencing. These proportions are log-transformed and the mean expression in each sample is subtracted from each value. This transformation converts the values into linearly independent values and the mean expression value of each sample is now 0. The difference between the within-condition distributions is determined by subtraction, giving a distribution of differences represented

below. The adjacent number represents the median difference. In step 3, distributions for the samples in each condition are merged into a joint distribution and the difference between conditions is determined by subtraction. Finally, as shown in step 4, the between-condition difference is divided by the larger of the within-condition differences. This results in a distribution that measures the difference between conditions scaled by the difference between samples in a condition. The method identifies genes that exhibit consistency of expression between samples within conditions and large differences between conditions. (See supporting text for a full description).

Figure S10

This figure is a complement to Figure 2: Circular representation of RNA-seq data for the *Lactobacillus iners* pan-transcriptome. RNA-seq reads were mapped to 12 available *L. iners*' genomes (listed in Table S3) after clustering redundant coding sequences by nucleotide identity (see Materials and Methods). Breaks in the circle separate blocks of contiguous coding sequences (CDS) ordered by scaffolding on the genome assemblies (Ring ①). The height of the plot represents how many of the 12 genomes the CDS occurs. The first contiguous block (dark blue) are the 954 CDS present in all genomes and are considered the “core” *L. iners* gene set. Ring ② shows the COG color-coded function of each CDS. Ring ③ The differential expression of each CDS between Non-BV and BV samples. The absolute height of the bar shows the fold-change (\log_2) in expression (positive for up BV, and negative for up non-BV). Bars are colored if significantly differential (red for BV and blue for non-BV). Regions of interest are labeled in the centre of the circle. Rings ④, ⑤, and ⑥ show a heatmap representation of the median CDS expression in BV, Non-BV, and *L. iners*

grown to mid-log phase in MRS broth respectively. Darker blue represents higher expression. Regions of interest noted in the results are marked and lettered as follows: A) Glycogen metabolism and uptake, B) Inolysin (CDC), C) PTS mannose-type, D) Glycerol metabolism, E) CRISPRs.

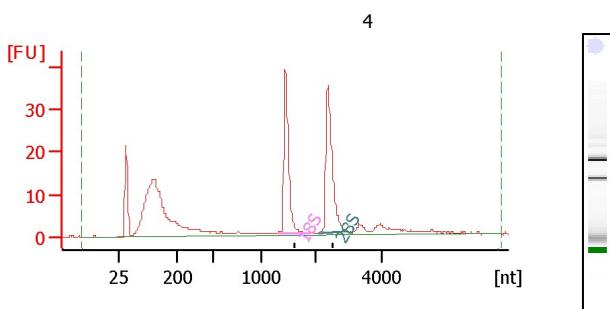
Tables S7 to S10 headers descriptions

refseq_id the identifier sequence for clustered CDS; **length** the refseq length in nt; **annotation** the refseq annotation from NCBI; **common_taxonomy** the lowest taxonomic level common to all sequences in the refseq cluster (represented by a single refseq ID); **cluster** the CD-HIT cluster number for the refseq; **no_seqs** the number of sequences in the CD-HIT cluster; **N4_total** the total number of reads mapped to this refseq (each sample in its own column); **subsys4 or kegg** the assigned SEED subsystem level 4 or kegg (KO number) function for the refseq; **expression**. the ALDEx-calculated median expression value for the refseq or function. This is relative to the average expression of all refseqs with the average expression being 0. A negative value represents expression below average, while a positive value is expression above average; **expression.within.N** or **B** The average median expression of the refseq within a condition (e.g. for N the average of **N4** and **N30** median expression); **expression.among.q50** the median expression across all samples (all conditions); **expression.sample**. median expression for the refseq/function in the indicated sample (**N4**, **N30**, **B27**, or **B31**); **difference.between.q50** median magnitude of change between conditions (median log2 fold change); **difference.within.q50** median maximum magnitude of change within conditions; **effect.q50** median relative difference; **criteria.significant** TRUE if **criteria.significance** is ≥ 0.01 (this is proportion of the

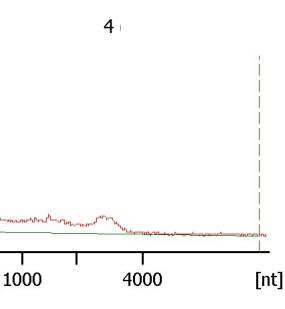
sample distributions that overlap between conditions); **criteria.meaningful** is TRUE if **criteria.meaning** is \geq the cutoff chosen (this is the relative difference. A value of 2 ensures that the between-condition is at least a 4-fold larger than the within-condition difference); **sig.both** TRUE if criteria.meaningful and criteria.significant are both TRUE. These represent our significantly differential genes/functions; **expression.D.q50** expression for a single sample of *L. iners* AB-1 grown in MRS broth

Figure S1**Total RNA****Enriched mRNA**

N4

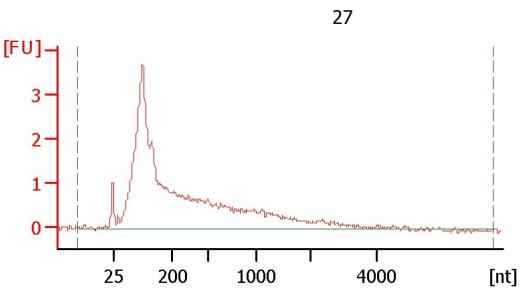
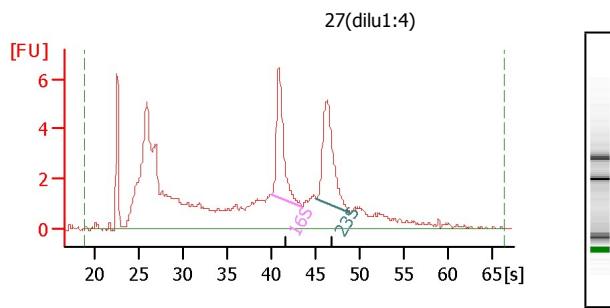
**Overall Results for sample**

| | |
|-----------------------------|--|
| RNA Area: | 318.5 |
| RNA Concentration: | 99 ng/ μ l |
| rRNA Ratio [23s / 16s]: | 1.2 |
| RNA Integrity Number (RIN): | 9.3 (B.02.07, Anomaly Threshold(s) manually adapted) |

**Overall Results for sample**

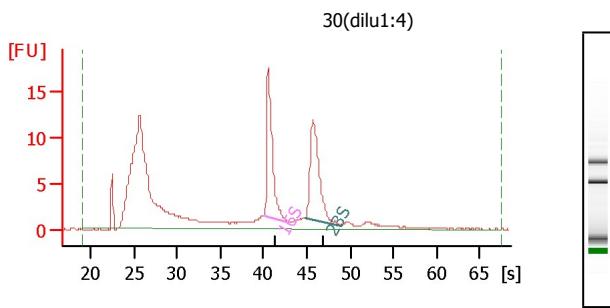
| | |
|-----------------------------|--|
| RNA Area: | 82.5 |
| RNA Concentration: | 213 ng/ μ l |
| rRNA Ratio [23s / 16s]: | 0.0 |
| RNA Integrity Number (RIN): | 2.3 (B.02.07, Anomaly Threshold(s) manually adapted) |

B27

**Overall Results for sample**

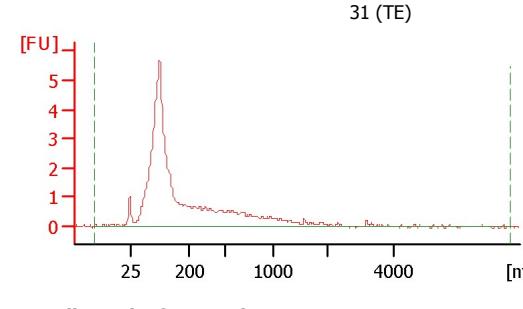
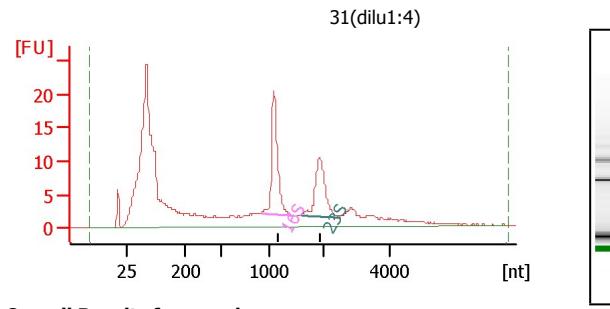
| | |
|-----------------------------|--|
| RNA Area: | 54.6 |
| RNA Concentration: | 141 ng/ μ l |
| rRNA Ratio [23s / 16s]: | 0.0 |
| RNA Integrity Number (RIN): | 2.2 (B.02.07, Anomaly Threshold(s) manually adapted) |

N30

**Overall Results for sample**

| | |
|-----------------------------|--|
| RNA Area: | 223.7 |
| RNA Concentration: | 186 ng/ μ l |
| rRNA Ratio [23s / 16s]: | 1.0 |
| RNA Integrity Number (RIN): | 8 (B.02.07, Anomaly Threshold(s) manually adapted) |

B31

**Overall Results for sample**

| | |
|-----------------------------|--|
| RNA Area: | 61.6 |
| RNA Concentration: | 159 ng/ μ l |
| rRNA Ratio [23s / 16s]: | 0.0 |
| RNA Integrity Number (RIN): | 2.3 (B.02.07, Anomaly Threshold(s) manually adapted) |

Overall Results for sample

| | |
|-----------------------------|--|
| RNA Area: | 362.6 |
| RNA Concentration: | 301 ng/ μ l |
| rRNA Ratio [23s / 16s]: | 0.7 |
| RNA Integrity Number (RIN): | 7.1 (B.02.07, Anomaly Threshold(s) manually adapted) |

Figure S2

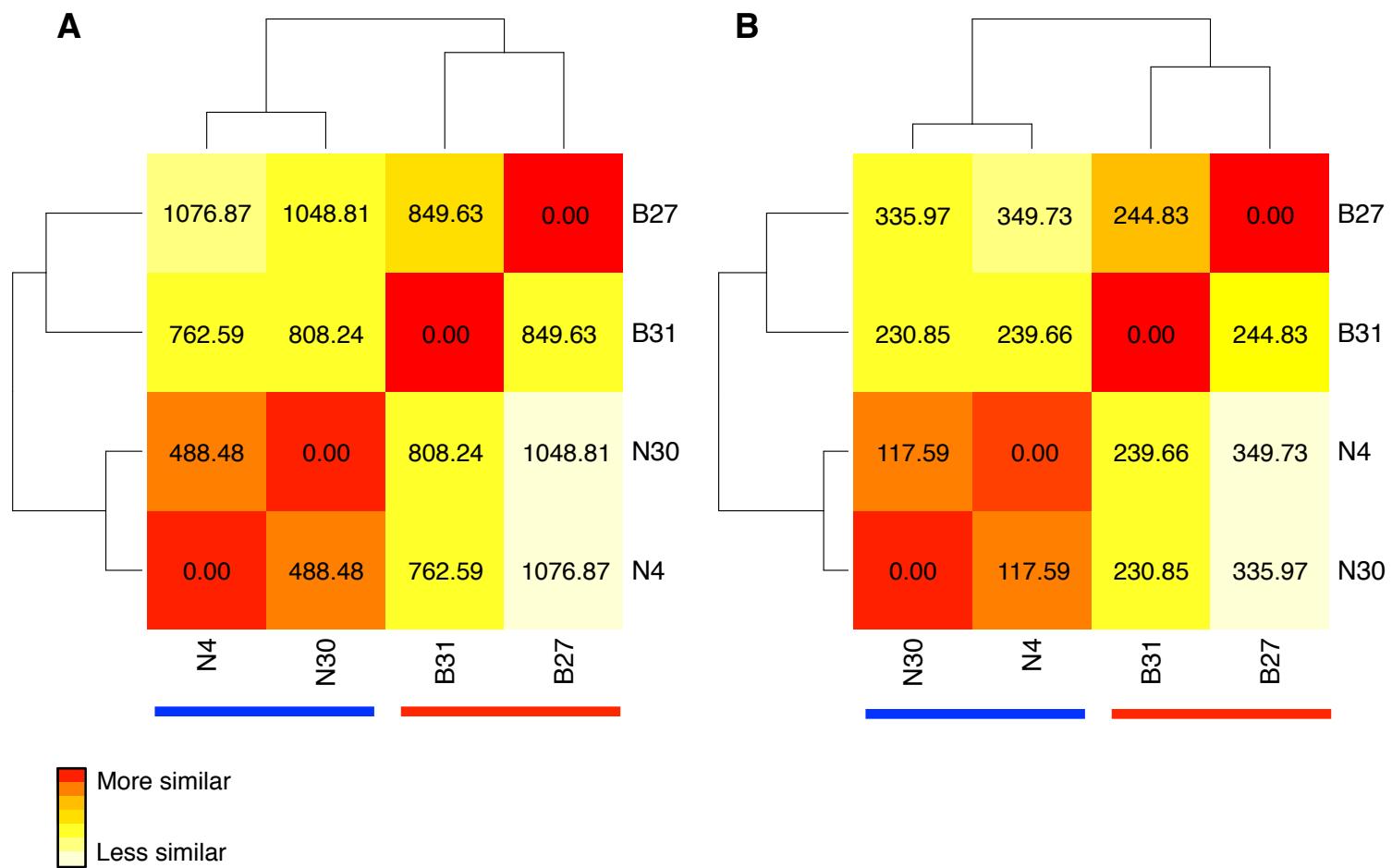


Figure S3A

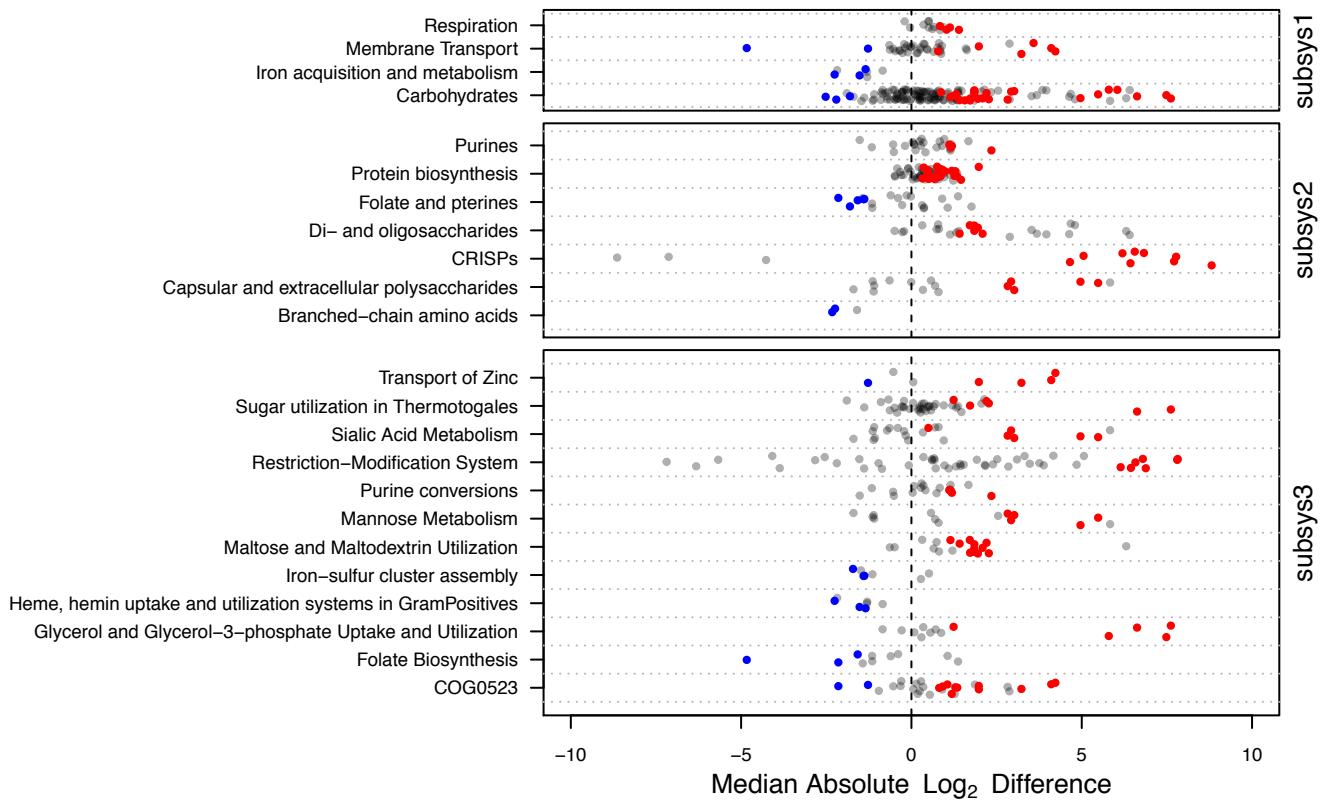


Figure S4

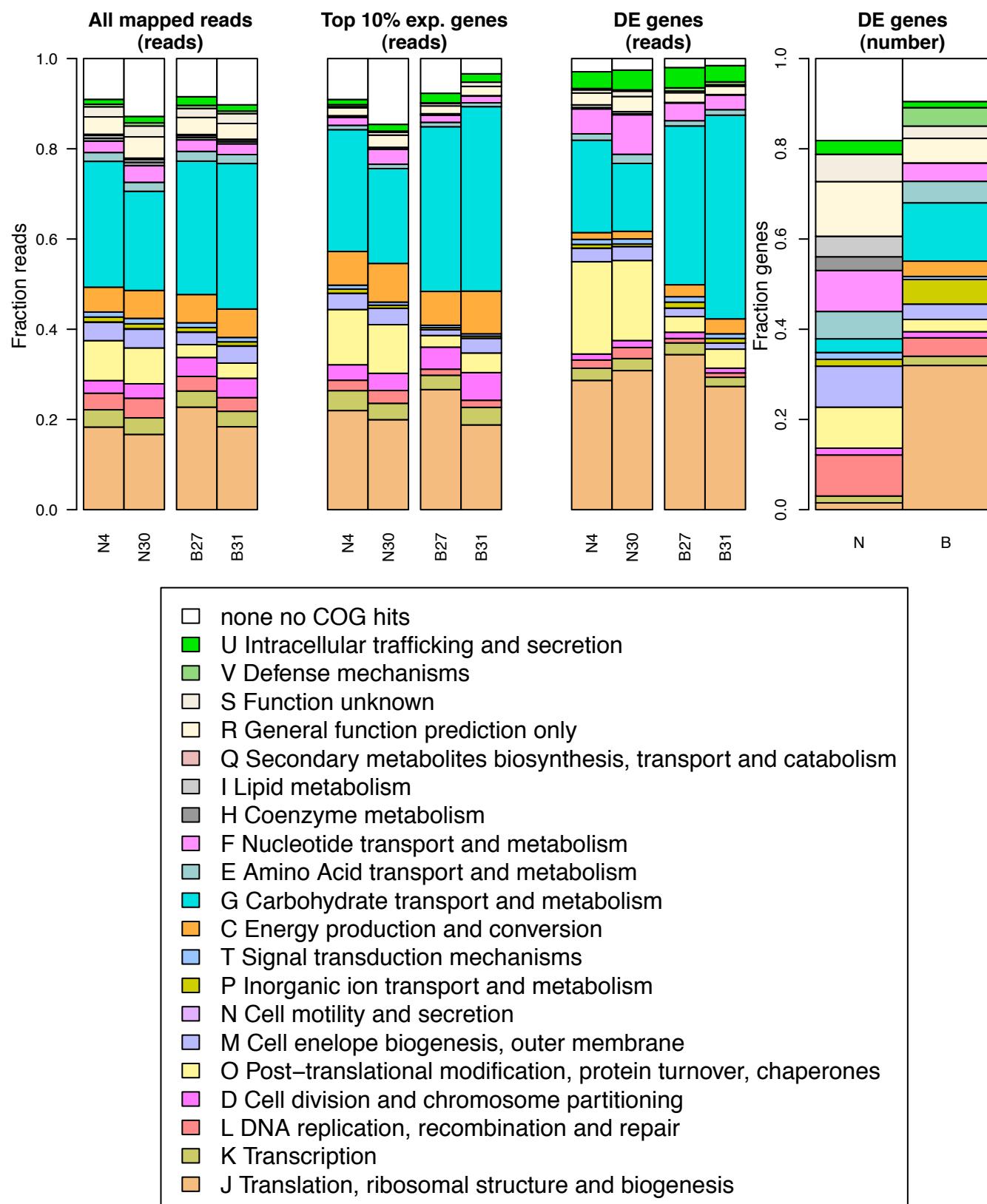


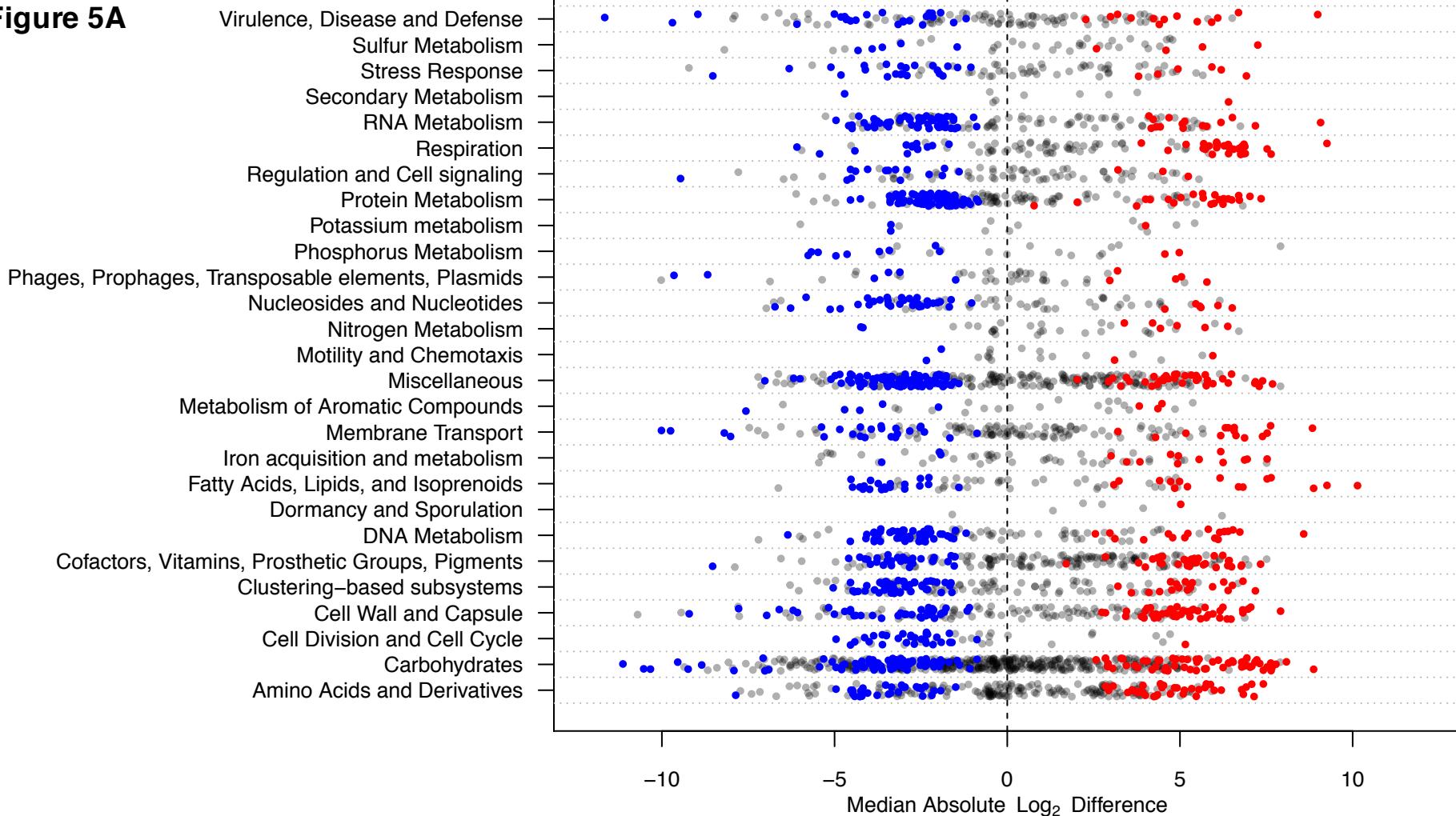
Figure 5A

Figure 5B

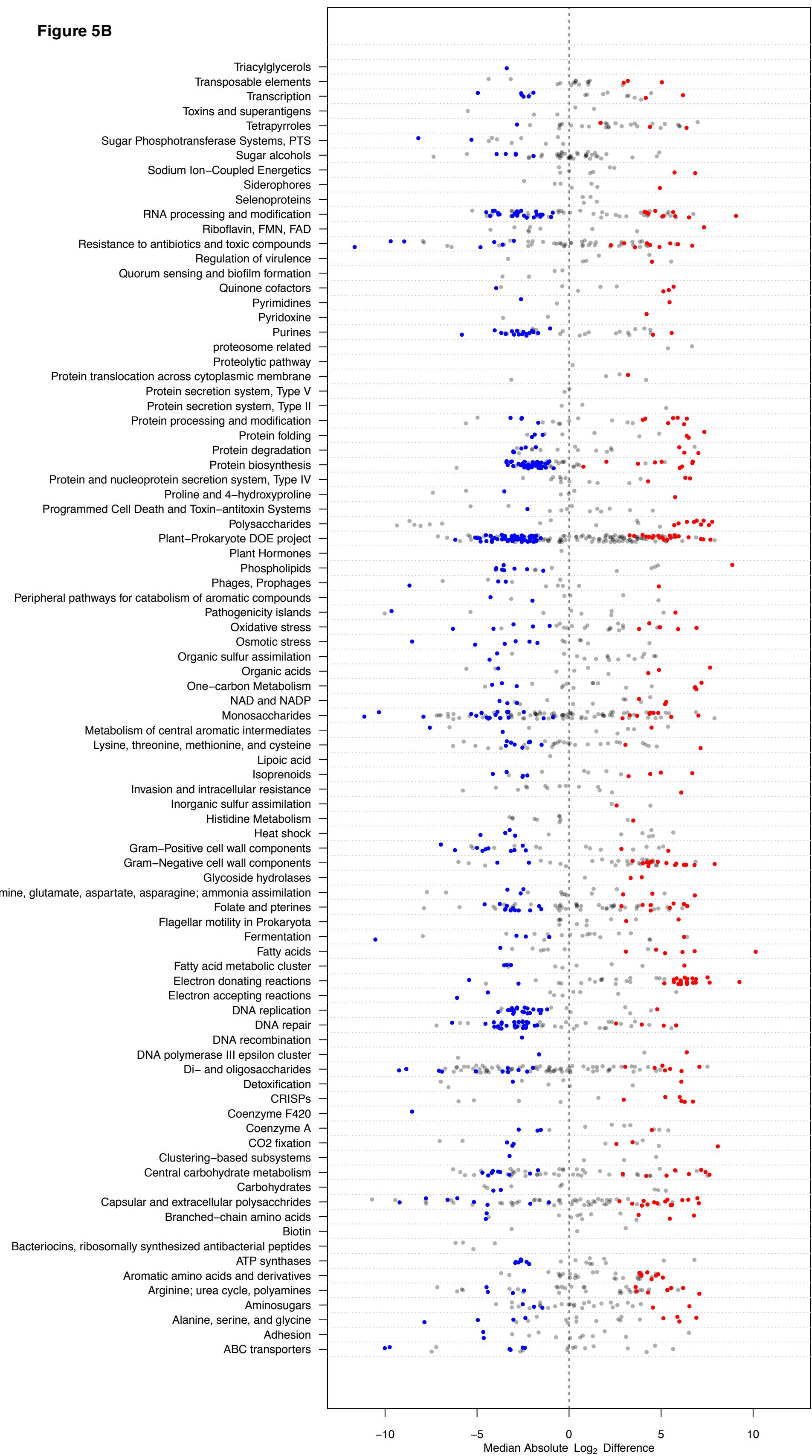


Figure 5



Figure S6A

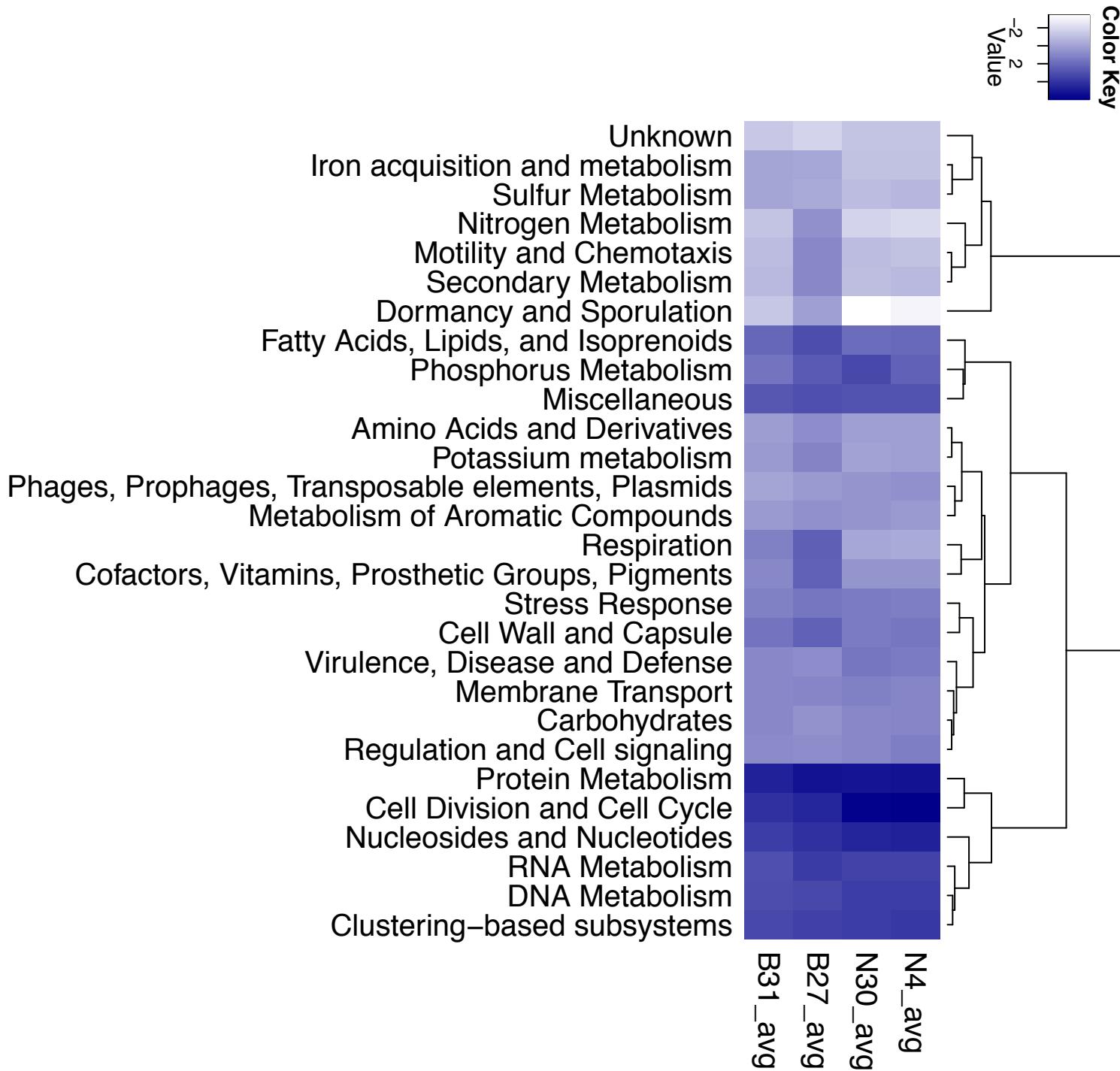


Figure S6B

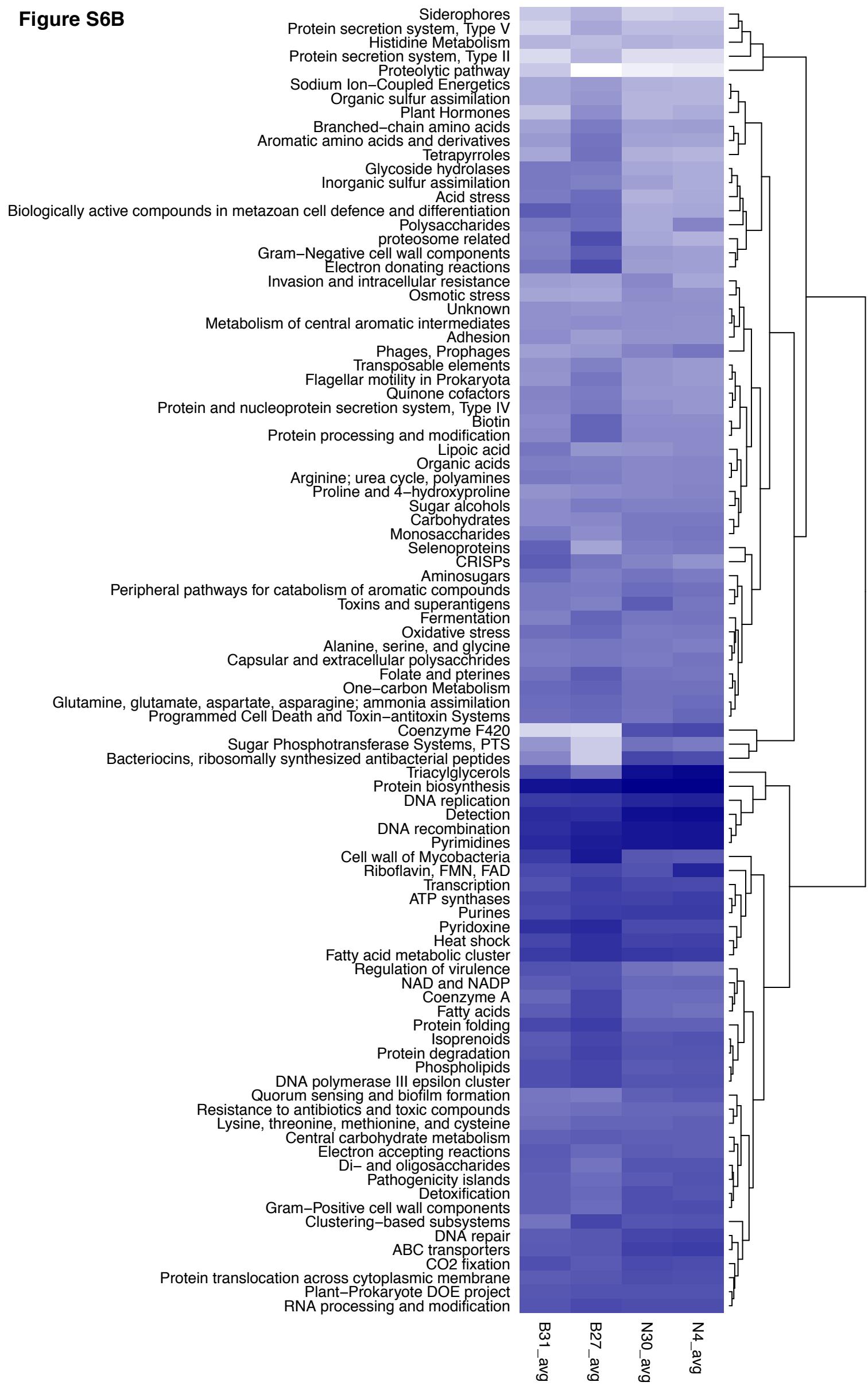


Figure S6C



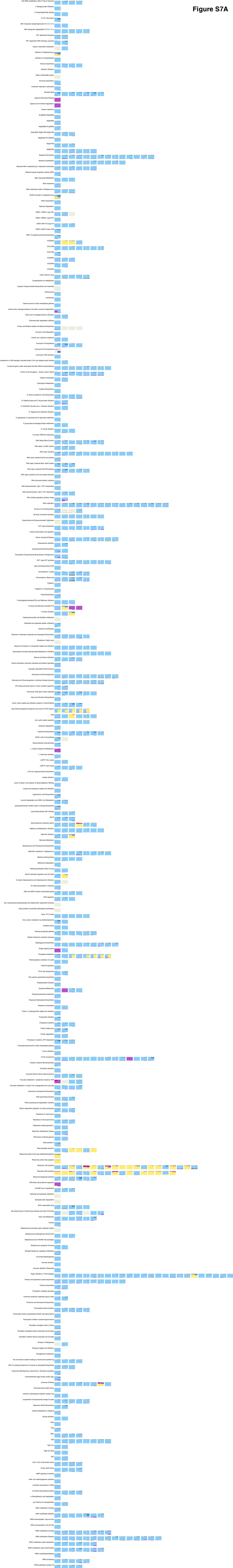
Figure S7A

Figure S7B

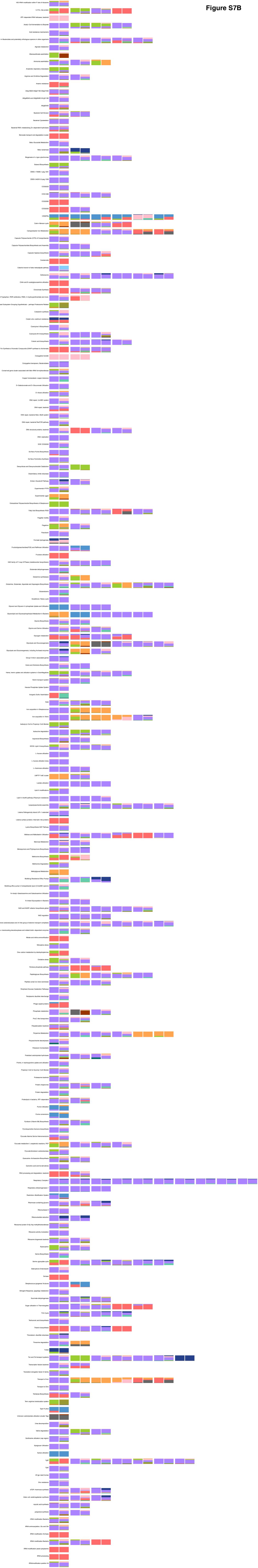


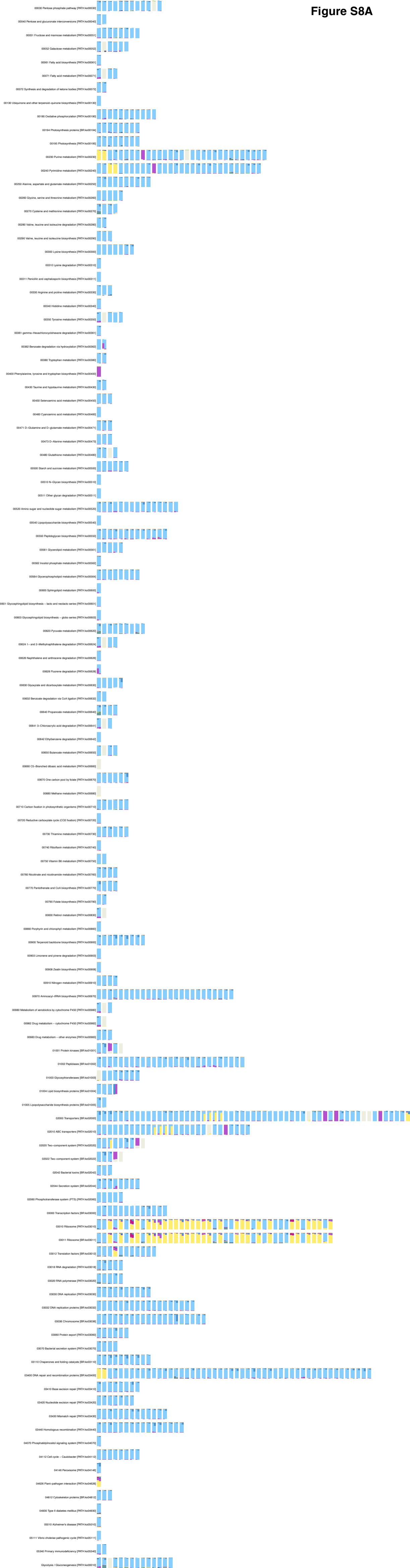
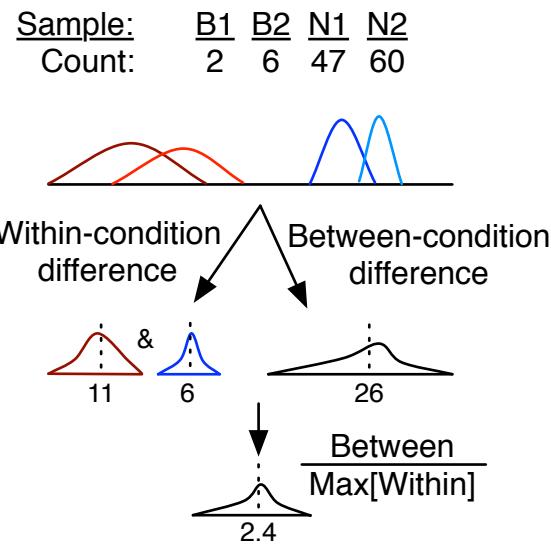
Figure S8A

Figure S8B

Figure S9



- 1) Read counts per function in each sample
- 2) Convert to proportional distributions
- 3) Calculate within and between condition distributions
- 4) Divide between-group difference by the largest within-group difference

Figure S10

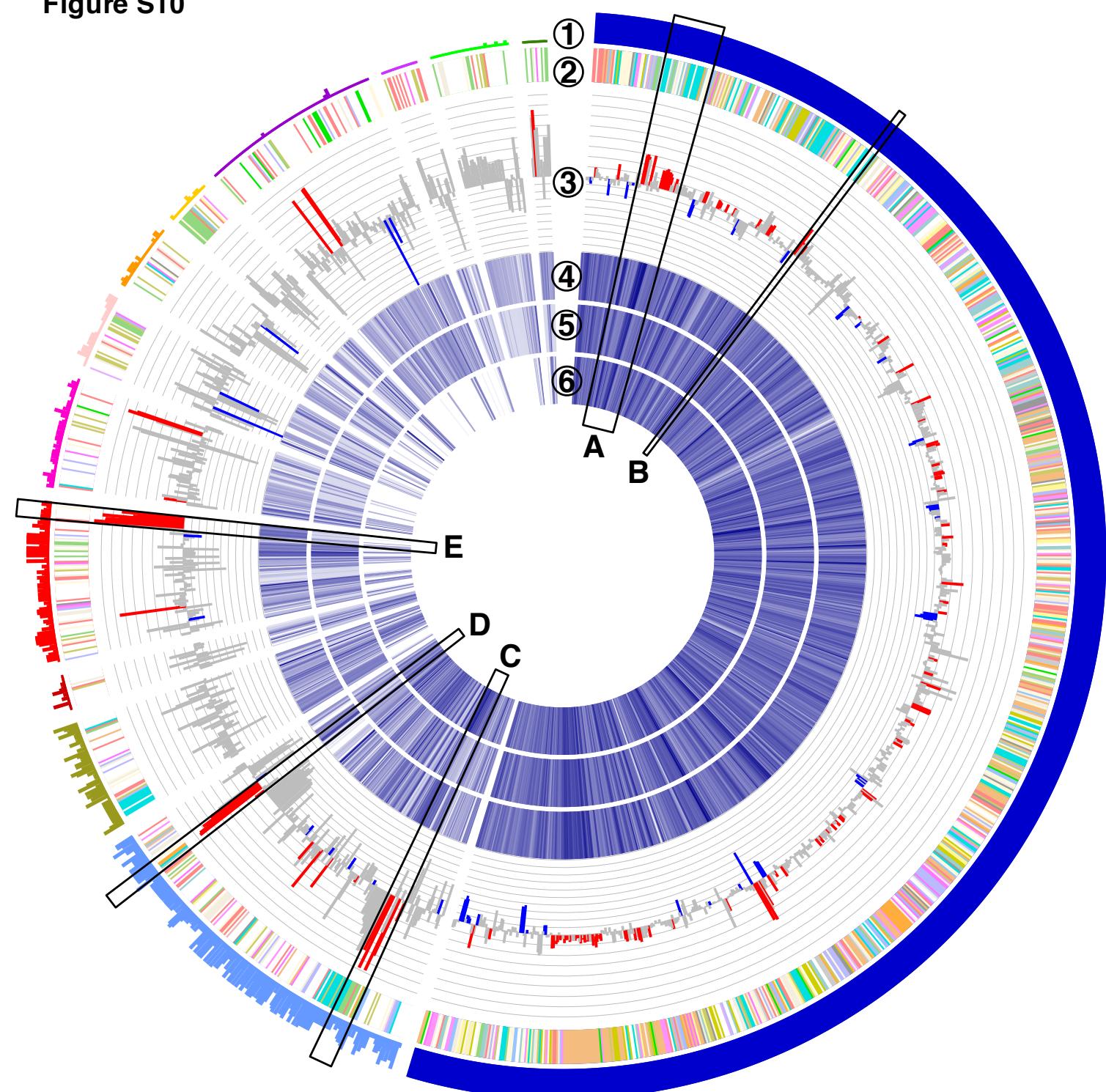


Table S1. Nugent scoring results

| Sample ID | Field No. | Lactobacillus Score | Gardnerella Score | Mobiluncus Score | Total Score | pH |
|-----------|-----------|---------------------|-------------------|------------------|-------------|--------|
| N4 | 1 | >30 | 0 | 0 | 2 | <= 4.5 |
| | 2 | >30 | 4 | 0 | | |
| | 3 | >30 | 2 | 0 | | |
| | 4 | >30 | 4 | 0 | | |
| | Average | >30 | 0 | 2.5 | 2 | |
| N30 | 1 | >30 | 1 | 0 | 2 | <= 4.5 |
| | 2 | >30 | 6 | 0 | | |
| | 3 | >30 | 0 | 0 | | |
| | 4 | >30 | 2 | 0 | | |
| | Average | >30 | 0 | 2.25 | 2 | |
| B27 | 1 | 0 | >30 | >30 | 10 | 5.5 |
| | 2 | 0 | >30 | >30 | | |
| | 3 | 0 | >30 | >30 | | |
| | 4 | 0 | >30 | >30 | | |
| | Average | 0 | 4 | >30 | 2 | |
| B31 | 1 | 16 | >30 | >30 | 7 | 5.0 |
| | 2 | 12 | >30 | >30 | | |
| | 3 | 12 | >30 | >30 | | |
| | 4 | 3 | >30 | >30 | | |
| | Average | 10.75 | 1 | >30 | 2 | |

Table S2. Sequence and mapping data.

| | | N4 | N30 | B27 | B31 | Total |
|--|------------------------------|---------------------|--------------------|--------------------|-------------------|---------------------------|
| Raw reads | | 47,634,967 | 46,943,184 | 48,832,687 | 53,655,422 | 197,066,260 |
| Reads mapped to human genome (excluded) | | 2,649,739 5.56% | 4,525,289 9.64% | 1,385,096 2.84% | 300,792 0.56% | 8,860,916 4.50% |
| Reads mapped to bacterial library | Total mapped | 20,556,493 | 14,690,153 | 11,981,948 | 12,556,674 | 59,785,268 |
| | Unique | 12,178,729 | 10,260,422 | 9,321,042 | 8,349,061 | 40,109,254 |
| | Non-unique | 8,377,764 | 4,429,731 | 2,660,906 | 4,207,613 | 19,676,014 |
| | Unique mapped to CDS* | 10,635,713 | 8,809,565 | 7,469,729 | 5,487,128 | 32,402,135 |
| No. of unique refseqs mapped | | 10,770 | 11,745 | 20,162 | 22,860 | 33,412[†] |
| No. refseqs assigned to KEGG | 18507 (55.4%) | 2056 unique KOs | | | | |
| No. refseqs assigned to SEED subsys4 | 22735 (68.0%) | 2772 unique subsys4 | | | | |
| No.refseqs assigned to COG | 25951 (77.7%) | 2300 unique COGs | | | | |

*These are reads mapping uniquely to coding sequences (CDS). These reads were used for differential expression analysis.

[†]This is the total number of unique reference sequence CDS (refseqs) mapped across all samples

Table S3. Read counts after mapping to curtated, non-redundant cpn60 database

| Taxon | 4N | 30N | 27B | 31B |
|-------------------------------|--------------|--------------|--------------|--------------|
| Lactobacillus_iners | 1060 | 6910 | 2759 | 5284 |
| Lactobacillus_crispatus | 14261 | 7010 | 0 | 707 |
| Gardnerella_vaginalis | 0 | 0 | 408 | 308 |
| Megasphaera_sp | 0 | 0 | 3202 | 0 |
| Prevotella_amnii | 15 | 12 | 34186 | 3113 |
| Prevotella_disiens | 0 | 0 | 0 | 918 |
| Prevotella_timonensis | 0 | 0 | 220 | 688 |
| Lactobacillus_jensenii | 222 | 194 | 0 | 17 |
| Lactobacillus_johnsonii | 22 | 0 | 0 | 0 |
| Lactobacillus_kefironofacieus | 190 | 0 | 0 | 0 |
| Lactobacillus_rhamnosus | 17 | 13 | 18 | 20 |
| Lactobacillus_acidophilus | 14 | 12 | 0 | 0 |
| Lactobacillus_ruminis | 0 | 0 | 46 | 0 |
| Alloscardovia_omnicolens | 0 | 0 | 15 | 0 |
| Atopobium_vaginae | 0 | 0 | 50 | 12 |
| Clostridium_genomosp_BVAB3 | 0 | 0 | 33 | 27 |
| Fusobacterium_sp | 0 | 0 | 23 | 0 |
| Peptoniphilus_lacrimalis | 0 | 0 | 0 | 48 |
| Porphyromonas_uenonis | 0 | 0 | 37 | 0 |
| Prevotella_bivia | 0 | 0 | 21 | 0 |
| Prevotella_buccalis | 0 | 0 | 42 | 14 |
| Prevotella_melaninogenica | 0 | 0 | 0 | 30 |
| Streptococcus_thermophilus | 145 | 102 | 81 | 94 |
| Vagococcus_fluvialis | 49 | 0 | 30 | 54 |
| total_reads | 15995 | 14253 | 41171 | 11334 |

Table S4. List of organisms used to build reference mapping library

| Organism name | GenBank or Refseq accession | No. of CDS |
|--|-----------------------------|------------|
| Actinomyces coleocanis DSM 15436 | ACFG000000000 | 1546 |
| Actinomyces urogenitalis DSM 15434 | ACFH000000000 | 2403 |
| Aerococcus viridans ATCC 11563 | ADNT000000000 | 1929 |
| Anaerococcus lactolyticus ATCC 51172 | ABYO000000000 | 2253 |
| Anaerococcus tetradius ATCC 35098 | ACGC000000000 | 2079 |
| Atopobium vaginae DSM 15829 | ACGK000000000 | 1271 |
| Atopobium vaginae PB189-T1-4 | AEDQ000000000 | 1217 |
| Bifidobacterium dentium ATCC 27679 | AEEQ000000000 | 2336 |
| Bifidobacterium dentium JCVIHMP022 | AEHJ000000000 | 2278 |
| Brevibacterium mcbrellneri ATCC 49030 | ADNU000000000 | 2432 |
| Chryseobacterium gleum F93, ATCC 35910 | ACKQ000000000 | 5289 |
| Clostridiales genomosp. BVAB3 UPII9-5 | CP001850 | 1526 |
| Corynebacterium aurimucosum ATCC 700975 | ACLH000000000 | 2614 |
| Corynebacterium genitalium ATCC 33030 | ACLJ000000000 | 2226 |
| Corynebacterium glucuronalyticum ATCC 51867 | ABYP000000000 | 2645 |
| Corynebacterium glucuronolyticum ATCC 51866 | ACHF000000000 | 2733 |
| Corynebacterium jeikeium ATCC 43734 | ACYW000000000 | 2224 |
| Corynebacterium lipophiloflavum DSM 44291 | ACHJ000000000 | 2371 |
| Corynebacterium pseudogenitalium ATCC 33035 | ABYQ000000000 | 2607 |
| Corynebacterium striatum ATCC 6940 | ACGE000000000 | 2677 |
| Dialister microaerophilus UPII-345-E | AENT000000000 | 1310 |
| Enterococcus faecalis ATCC 29200 | ACHK000000000 | 2955 |
| Enterococcus faecalis HH22 | ACIX000000000 | 3227 |
| Enterococcus faecalis TX 0312 | AECB000000000 | 2789 |
| Enterococcus faecalis TX 0635 | AEBZ000000000 | 3263 |
| Enterococcus faecalis TX0855 | AEBV000000000 | 3032 |
| Eremococcus coleocola ACS-139-V-Col8 | AENN000000000 | 1720 |
| Escherichia coli 83972 | ACGN000000000 | 5255 |
| Finegoldia magna ACS-171-V-Col3 | AECM000000000 | 1740 |
| Finegoldia magna ATCC 29328 | NC_010371 | 182 |
| Finegoldia magna ATCC 29328 | NC_010376 | 1631 |
| Finegoldia magna ATCC 53516 | ACHM000000000 | 1838 |
| Finegoldia magna BVS033A4 | AEDP000000000 | 1904 |
| Fusobacterium nucleatum nucleatum ATCC 23726 | ADVK000000000 | 2126 |
| Gardnerella vaginalis ATCC 14019 | NC_014644 | 1365 |
| Gardnerella vaginalis 409-05 | NC_013721 | 1261 |
| Gardnerella vaginalis AMD | ADAM000000000 | 1313 |
| Lactobacillus acidophilus NCFM | NC_006814 | 1864 |
| Lactobacillus crispatus 214-1 | ADGR000000000 | 2163 |
| Lactobacillus crispatus JV-V01 | ACKR000000000 | 2209 |
| Lactobacillus crispatus MV-3A-US | ACQC000000000 | 2330 |
| Lactobacillus crispatus ST1 | NC_014106 | 2024 |
| Lactobacillus delbrueckii bulgaricus PB2003/044-T3-4 | AEAT000000000 | 1909 |
| Lactobacillus fermentum IFO | NC_010610 | 1843 |
| Lactobacillus gasseri 202-4 | ACOZ000000000 | 1773 |
| Lactobacillus gasseri 224-1 | ADFT000000000 | 2252 |
| Lactobacillus gasseri JV-V03 | ACGO000000000 | 1977 |
| Lactobacillus gasseri MV-22 | ABWH000000000 | 1917 |
| Lactobacillus iners AB-1 | ADHG000000000 | 1190 |
| Lactobacillus iners ATCC 55195 | AEPX000000000 | 1144 |
| Lactobacillus iners DSM 13335 | ACLN000000000 | 1214 |
| Lactobacillus iners LactinV 01V1-a | AEHQ000000000 | 1527 |
| Lactobacillus iners LactinV 03V1-b | AEHP000000000 | 1459 |
| Lactobacillus iners LactinV 09V1-c | AEHO000000000 | 1361 |
| Lactobacillus iners LactinV 11V1-d | AEHN000000000 | 1338 |

| Organism name | GenBank or Refseq accession | No. of CDS |
|---|------------------------------------|-------------------|
| Lactobacillus iners LEAF 2052A-d | AEKI00000000 | 1256 |
| Lactobacillus iners LEAF 2053A-b | AEKH00000000 | 1277 |
| Lactobacillus iners LEAF 2062A-h1 | AEKJ00000000 | 1265 |
| Lactobacillus iners LEAF 3008A-a | AEKK00000000 | 1210 |
| Lactobacillus iners SPIN 2503V10-D | AEHR00000000 | 1273 |
| Lactobacillus jensenii 1153 | ABWG00000000 | 1395 |
| Lactobacillus jensenii 208-1 | ADEX00000000 | 3182 |
| Lactobacillus jensenii 269-3 | ACOY00000000 | 1575 |
| Lactobacillus jensenii JV-V16 | ACGQ00000000 | 1450 |
| Lactobacillus johnsonii ATCC 33200 | ACGR00000000 | 1838 |
| Lactobacillus johnsonii FI9785 | NC_012552 | 2 |
| Lactobacillus johnsonii FI9785 | NC_013504 | 1710 |
| Lactobacillus johnsonii FI9785 | NC_013505 | 23 |
| Lactobacillus johnsonii FI9785 | NC_005362 | 1821 |
| Lactobacillus oris PB013-T2-3 | AEKL00000000 | 2038 |
| Lactobacillus reuteri_JCM 1112 | NC_010609 | 1820 |
| Lactobacillus rhamnosus_Lc 705 | NC_013199 | 2878 |
| Lactobacillus rhamnosus_Lc 705 | NC_013200 | 79 |
| Lactobacillus salivarius ACS-116-V-Col5a | AEBA00000000 | 2121 |
| Lactobacillus vaginalis ATCC 49540 | ACGV00000000 | 1870 |
| Leptotrichia buccalis C 1013 | NC_013192 | 2220 |
| Megasphaera genomosp. type_1 28L | ADGP00000000 | 1610 |
| Mobiluncus curtisii ATCC 43063 | NC_014246 | 1909 |
| Mobiluncus curtisii ATCC 51333 | AEPY00000000 | 1855 |
| Mobiluncus curtisii curtisii ATCC 35241 | AAEE00000000 | 1894 |
| Mobiluncus curtisii holmesii ATCC 35242 | AEPE00000000 | 1829 |
| Mobiluncus mulieris 28-1 | ADBR00000000 | 2321 |
| Mobiluncus mulieris ATCC 35239 | AEET00000000 | 2380 |
| Mobiluncus mulieris ATCC 35243 | ACKW00000000 | 2300 |
| Mobiluncus mulieris FB024-16 | AEGV00000000 | 2113 |
| Mycobacterium parascrofulaceum ATCC BAA-614 | ADNV00000000 | 6456 |
| Peptoniphilus duerdenii ATCC BAA-1640 | AEEH00000000 | 1988 |
| Peptoniphilus lacrimalis 315-B | ADDO00000000 | 1589 |
| Porphyromonas asaccharolytica PR426713P-I | AENO00000000 | 1655 |
| Porphyromonas uenonis 60-3 | ACLR00000000 | 1977 |
| Prevotella amnii CRIS 21A-A | ADFQ00000000 | 2025 |
| Prevotella bivia JCVIHMPO10 | ADFO00000000 | 2041 |
| Prevotella buccalis ATCC 35310 | ADEG00000000 | 2456 |
| Prevotella disiens FB035-09AN | AEDO00000000 | 2621 |
| Prevotella melaninogenica ATCC 25845 | NC_014370 | 1338 |
| Prevotella melaninogenica ATCC 25845 | NC_014371 | 958 |
| Prevotella oralis ATCC 33269 | AEPE00000000 | 2488 |
| Prevotella timonensis CRIS 5C-B1 | ADEF00000000 | 2202 |
| Proteus mirabilis ATCC 29906 | ACLE00000000 | 3812 |
| Roseomonas cervicalis ATCC 49957 | ADVL00000000 | 4778 |
| Sphingobacterium spiritivorum ATCC 33300 | ACHB00000000 | 4925 |
| Sphingobacterium spiritivorum ATCC 33861 | ACHA00000000 | 4471 |
| Staphylococcus aureus aureus MN8 | ACJA00000000 | 2833 |
| Staphylococcus aureus COL | NC_002951 | 2612 |
| Staphylococcus aureus COL | NC_006629 | 3 |
| Streptococcus agalactiae 2603V | NC_004116 | 2124 |
| Streptococcus bovis ATCC 700338 | AEEL00000000 | 2088 |
| Streptococcus pseudoporcinus SPIN 20026 | AENS00000000 | 2030 |
| Treponema phagedenis F0421 | AEFH00000000 | 3146 |
| Veillonella atypica ACS-049-V-Sch6 | AEDR00000000 | 1840 |
| Total | | 230031 |

Table S5. Read coverage of *Lactobacillus iners* coding sequences (CDS)

Average number of CDS* 1293

Average CDS length (nt) 970

Product **1254210**

| | N4 | N30 | N27 | N31 |
|---|-------------|--------------|--------------|---------------|
| Total mapped reads to <i>L. iners</i>' CDS | 122332 | 617960 | 1928856 | 2931047 |
| x 50 bp per read | 6116600 | 30898000 | 96442800 | 146552350 |
| Approx. CDS coverage | 4.88 | 24.64 | 76.90 | 116.85 |

*See Table S4

Table S6. Summary of reads mapped to CRISPR spacer sequences

| CRISPR spacer sequence ID | N4 | | N30 | | B27 | | B31 | |
|---|---------------|-------------------------|---------------|-------------------------|--------------|-------------------------|--------------|-------------------------|
| | No. reads | % of total mapped reads | No. reads | % of total mapped reads | No. reads | % of total mapped reads | No. reads | % of total mapped reads |
| Icl 1713 ref NC_007716 Aster yellows witches'-broom phytoplasma AYWb 277436-277525 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 1 | 0.00 |
| Icl 1714 ref NC_007716 Aster yellows witches'-broom phytoplasma AYWb 277342-277417 | 58 | 0.03 | 109 | 0.07 | 15 | 0.07 | 55 | 0.06 |
| Icl 1716 ref NC_007716 Aster yellows witches'-broom phytoplasma AYWb 277167-277226 | 0 | 0.00 | 0 | 0.00 | 1 | 0.00 | 0 | 0.00 |
| Icl 2498 ref NC_011891 Anaeromyxobacter dehalogenans 2CP-1 3151161-3151315 | 0 | 0.00 | 1 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Icl 3269 ref NC_009674 Bacillus cereus subsp. cytotoxis NVH 391-98 3571402-3571478 | 8 | 0.00 | 18 | 0.01 | 3 | 0.01 | 7 | 0.01 |
| Icl 3270 ref NC_009674 Bacillus cereus subsp. cytotoxis NVH 391-98 3571303-3571377 | 10 | 0.00 | 95 | 0.06 | 228 | 1.06 | 779 | 0.81 |
| Icl 3272 ref NC_009674 Bacillus cereus subsp. cytotoxis NVH 391-98 3571141-3571198 | 21 | 0.01 | 250 | 0.16 | 95 | 0.44 | 356 | 0.37 |
| Icl 10652 ref NC_008593 Clostridium novyi NT 1866066-1866100 | 2 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Icl 13253 ref NC_012785 Kosmotoga olearia TBF 19.5.1 723753-723786 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 1 | 0.00 |
| Icl 13330 ref NC_010080 Lactobacillus helveticus DPC 4571 1587946-1587980 | 15 | 0.01 | 17 | 0.01 | 0 | 0.00 | 0 | 0.00 |
| Icl 13337 ref NC_010080 Lactobacillus helveticus DPC 4571 1695949-1696043 | 101829 | 45.21 | 63058 | 41.28 | 14304 | 66.60 | 56128 | 58.03 |
| Icl 13338 ref NC_010080 Lactobacillus helveticus DPC 4571 1695809-1695937 | 90804 | 40.31 | 63592 | 41.63 | 6194 | 28.84 | 31881 | 32.96 |
| Icl 13339 ref NC_010080 Lactobacillus helveticus DPC 4571 1695675-1695797 | 27975 | 12.42 | 24092 | 15.77 | 15 | 0.07 | 5400 | 5.58 |
| Icl 13340 ref NC_010080 Lactobacillus helveticus DPC 4571 1695542-1695663 | 2698 | 1.20 | 1254 | 0.82 | 441 | 2.05 | 1408 | 1.46 |
| Icl 15149 ref NC_006360 Mycoplasma hyopneumoniae 232 338057-338117 | 0 | 0.00 | 0 | 0.00 | 45 | 0.21 | 33 | 0.03 |
| Icl 15151 ref NC_006360 Mycoplasma hyopneumoniae 232 337855-337962 | 1 | 0.00 | 1 | 0.00 | 22 | 0.10 | 178 | 0.18 |
| Icl 16408 ref NC_013316 Clostridium difficile R20291 chromosome 1292772-1292809 | 1 | 0.00 | 8 | 0.01 | 0 | 0.00 | 0 | 0.00 |
| Icl 16502 ref NC_010376 Finegoldia magna ATCC 29328 chromosome 197417-197654 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 3 | 0.00 |
| Icl 16505 ref NC_010376 Finegoldia magna ATCC 29328 chromosome 196644-196855 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 4 | 0.00 |
| Icl 17873 ref NC_009454 Pelotomaculum thermopropionicum SI chromosome 1998702-1998738 | 0 | 0.00 | 0 | 0.00 | 1 | 0.00 | 0 | 0.00 |
| Icl 18404 ref NC_006814 Lactobacillus acidophilus NCFM chromosome 1541067-1541098 | 13 | 0.01 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Icl 18406 ref NC_006814 Lactobacillus acidophilus NCFM chromosome 1542898-1542929 | 37 | 0.02 | 30 | 0.02 | 33 | 0.15 | 122 | 0.13 |
| Icl 20031 ref NC_012032 Chloroflexus sp. Y-400-fl chromosome 4548942-4548978 | 0 | 0.00 | 0 | 0.00 | 8 | 0.04 | 17 | 0.02 |
| Icl 20808 ref NC_013511 Mycoplasma hominis 259090-259162 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 1 | 0.00 |
| Icl 25524 ref NC_009523 Roseiflexus sp. RS-1 chromosome 3157153-3157189 | 0 | 0.00 | 3 | 0.00 | 0 | 0.00 | 1 | 0.00 |
| Icl 25767 ref NC_014752 Neisseria lactamica ST-640 1672927-1672964 | 1770 | 0.79 | 229 | 0.15 | 66 | 0.31 | 310 | 0.32 |
| Icl 26144 ref NC_014484 Spirochaeta thermophila DSM 6192 chromosome 461133-461170 | 0 | 0.00 | 0 | 0.00 | 1 | 0.00 | 0 | 0.00 |
| Icl 31970 ref NC_000919 Treponema pallidum subsp. pallidum str. Nichols chromosome 77273-77566 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 15 | 0.02 |
| Icl 32709 ref NC_010175 Chloroflexus aurantiacus J-10-fl chromosome 4538893-4538930 | 0 | 0.00 | 0 | 0.00 | 5 | 0.02 | 17 | 0.02 |
| Icl 33777 ref NC_010741 Treponema pallidum subsp. pallidum SS14 77270-77563 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 10 | 0.01 |
| Icl 39861 ref NC_015499 Thermodesulfobium narugense DSM 14796 chromosome 1770147-1770185 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 1 | 0.00 |
| Icl 35999 ref NC_013715 Rothia mucilaginosa DY-18 chromosome 1380875-1381203 | 0 | 0.00 | 1 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| Icl 40601 ref NC_009437 Caldicellulosiruptor saccharolyticus DSM 8903 chromosome 357323-357358 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 2 | 0.00 |
| Total reads mapped | 225242 | | 152758 | | 21477 | | 96730 | |